Advances in Medical Sciences 64 (2019) 181–188

Contents lists available at ScienceDirect

Advances in Medical Sciences

journal homepage: www.elsevier.com/locate/advms

Original research article

Ensembling rules in automatic analysis of pressure on plantar surface in children with pes planovalgus

Marcin Derlatka[a], Mikhail Ihnatouski[b], Marek Jałbrzykowski[a,*], Vladimir Lashkovski[c], Łukasz Minarowski[d]

[a] Faculty of Mechanical Engineering, Bialystok University of Technology, Bialystok, Poland
[b] Scientific and Research Department, Yanka Kupala State University of Grodno, Grodno, Belarus
[c] Department of Traumatology, Orthopedics and Field Surgery, Grodno State Medical University, Grodno, Belarus
[d] 2nd Department of Lung Diseases and Tuberculosis, Medical University of Bialystok, Bialystok, Poland

ARTICLE INFO

Keywords:
Human gait analysis
Decision trees
Ensemble rules
Pes planovalgus
Pressure distribution

ABSTRACT

Purpose: This paper presents a method of ensembling rules obtained through induction of several basic types of decision trees.
Material and methods: The proposed method uses rules generated by means of well-known decision trees: CART, CHAID, exhaustive CHAID and C4.5. The method was tested on data describing pressure distribution under foot during gait in children with pes planovalgus (PV) and typical foot. Children with pes planovalgus underwent surgical intervention and were re-examined. Overall, 316 gait cycles have been used in analysis.
Results: The obtained results consist of a set of rules for all considered cases and show that the proposed method may be a useful tool of gait analysis on the basis of parameters that have a physical interpretation.
Conclusions: The presented method for mining rules useful in this respect may be successfully used by persons with a typically medical knowledge and could improve the understanding of the human gait phenomenon. There is obviously no reason why this method could not be used in the case of other data as well.

1. Introduction

Foot disorders are one of the most common health problems in the schoolchildren group [1]. Pes planovalgus (PV) is the most common foot disease in which the foot has a small, longitudinal arch during loading, resulting from a failure of the muscular-ligament system. Moreover, in PV the heel bone is in pronation, which results from the foot being more flattened. The results of PV are foot deformation and pain, for this reason, children should be correctly diagnosed and start therapy as soon as possible.

Modern technologies make is possible to use many different methods to analyze the state of the human foot. Among the most common there are: radiological examination [2], motion capture systems equipped with force plates [3,4], and the pedobarograph [5,6]. Special attention should be paid to plantar pressure measurement systems. They are becoming important devices used in gait analysis, especially in foot disorders, as they provide relatively complete information about the working of the foot both in static and dynamic conditions [7–10]. This enables to make a deeper analysis of the state of the examined feet and elaborate in a more accurate way on the course of the rehabilitation process.

Unfortunately, the measurement data obtained from systems of this type are time series that prove at least problematic in direct interpretation (Fig. 1). This results in parametrization of measurement data in a way that would enable their fast and effective analysis. It is equally important to perform the analysis on the basis of indicators that not only have a physical interpretation, but are also resistant to measurement errors and outliers. Analysis of this kind is enabled by data mining, i.e. a set of methods that makes it possible to manage huge and multidimensional sets of measurement data and perform its fast and efficient analysis as well as find new, sometimes unexpected, connections between various parameters. Data mining methods are also used in the field of biomedical and human gait analysis. Their popularity results from their breaking the limitations of manual evaluation of gait-related data [11,12]. One of the most promising data mining techniques are decision trees.

Decision trees enable extracting the knowledge hidden in the data and presenting it in a very vivid way. They provide very simple conditions in the tree nodes and lead to a conclusion (class) on the lowest level of the tree. A very important aspect is that the results are easy to

* Corresponding author at: Faculty of Mechanical Engineering, Bialystok University of Technology, Wiejska 45C, 15-351 Bialystok, Poland.
E-mail address: m.jalbrzykowski@pb.edu.pl (M. Jałbrzykowski).

https://doi.org/10.1016/j.advms.2018.08.009
Received 5 March 2018; Accepted 31 August 2018
Available online 02 February 2019

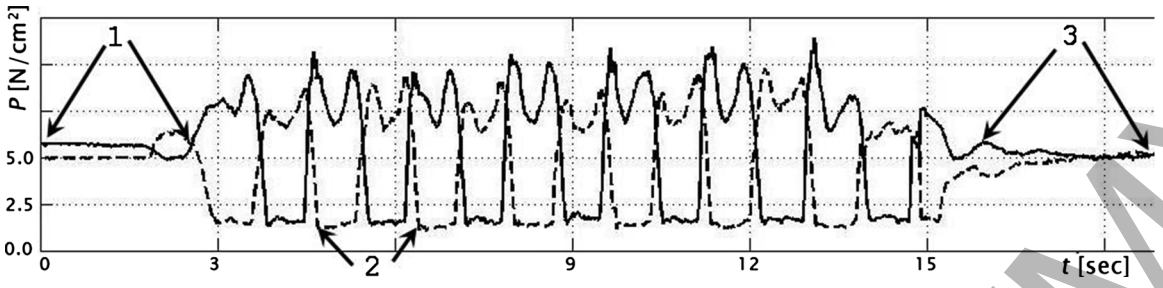

Fig. 1. Diagram of the total pressure (*P*) under the feet (solid line – left, dot line – right): 1 and 3 period of immobility (static); 2 walk (dynamic).

interpret and could be used by the staff with no mathematical or engineering background. Decision trees have already been successfully used in human gait analysis in clinical applications [10,13,14]. It is worth noting that decision trees make it very easy to read rules in the following form:

If (condition1) *AND* (condition2) … *AND* (conditionN) then *diagnosis*

The value of such rules is higher the higher the percentage of correct assignments of input data to classes (e.g. pathology types).

Currently, one of the most widely-known methods of improvement of classification is combining classifiers [15]. In the case of decision trees, the most commonly proposed methods focus on determining a linear combination of individual rules. The rules are presented as a weighted sum of all the tree outputs [16]:

$$\sum_{l=1}^{L} w_l \bullet A_l > c$$

where: w_l – a weight calculated during creation of rule, A_l – an attribute which can be physically interpreted, c - a threshold value, L – the number of augends in a single condition.

The method, however, leads to the creation of rules that are extremely non-intuitive in analysis. Recently, a few methods which enable extracting the rules based on ensemble decision trees have been elaborated [17,18]. Moreover, there exist several methods which allow to generate a set of rules without creating a decision tree. Among those methods there are: JRiP [19], Decision Table, PART [16] or MODLEM [20]. However, all rules obtained from decision trees as well as obtained by means of method to induce set of rules have the same disadvantage - assigning of input data to one or several possible classes even if the analyzed data does not display any similarity with the other examples. In medical applications such course of action is ill-considered insomuch as a case that is atypical, representing a dysfunction other than those under consideration or requiring a different approach, should not be classified as one of the possible classes.

This study deals with the problem of creating new reliable decision rules sourced from heterogeneous decision trees.

2. Materials and method

The literature provides a number of types of decision trees differing in their methods of creation (induction) on the basis of a data set. Among the most popular decision trees there are: CART, CHAID, exhaustive CHAID, C4.5, and random trees.

2.1. Decision trees

2.1.1. CART

Classification And Regression Tress (CART) are binary trees (each node has no more than 2 children) with one-dimensional splits. The condition in the tree node is created by checking all possible splits, in points that are the centers of the segments between subsequent sorted values x_j and x_{j+1}. The optimum split is one that splits input data into possibly the most homogeneous subsets. Homogeneity assessment $i(t)$ after split can be performed with the use of, for example, Gini's index:

$$i(t) = 1 - \sum_{k=1}^{K} p_k^2$$

where:

- p_k – probability of occurrence of k class elements after split
- K – the number of all classes.

2.1.2. CHAID

Chi-square Automatic Interaction Detector (CHAID algorithm creates trees with nodes that can have more than two children. Initially CHAID trees were used when the input data was qualitative. Modification of the original algorithm enables to use them for quantitative data as well after their conversion into qualitative data (split into approximately equal numbers of observations). CHAID algorithm views the predictors one by one and searches for a pair of categories for each of them differing in at least one dependent variable. The analysis is performed by conducting the Chi-square test. If for a given pair of categories, the test does not yield a statistically significant difference for the pair, the program joins the categories and repeats the step. If p-value is statistically significant (lower than the corresponding p level value for the pair), the program calculates the Bonferroni-corrected p-value for the set of predictor categories. In the next step, the program chooses the predictor with the lowest value of p level (corrected), i.e. the predictor that yields the most significant split. If the p level value (Bonferroni-corrected) for each predictor is lower than the p level for the split, then subsequent splits are not performed and the node is a tree leaf.

2.1.3. Exhaustive CHAID

Exhaustive CHAID is a modified CHAID that enables a more accurate analysis of all the possible splits for each predictor. This results in much longer calculation times.

2.1.4. C4.5

C4.5 tree algorithm chooses, for each tree node, an attribute that most effectively splits a set of objects (i.e. objects within a single class must dominate in each of the subtrees of a given node). Subtree entropy is thus minimized, described with the following equation:

$$i(t) = -\sum_{k=1}^{K} p_k \cdot \log p_k$$

The gain ratio resulting from choosing a given attribute is:

$$\text{gain(A)} = \text{i(t)-E(A)}$$

where:

E(A) is the weighted average subnode entropy, where the weighs are the ratios between the number of subnode elements and the number of elements in the whole node.

2.1.5. Random Forest

Random Forest is a set of many relatively simple decision trees. In

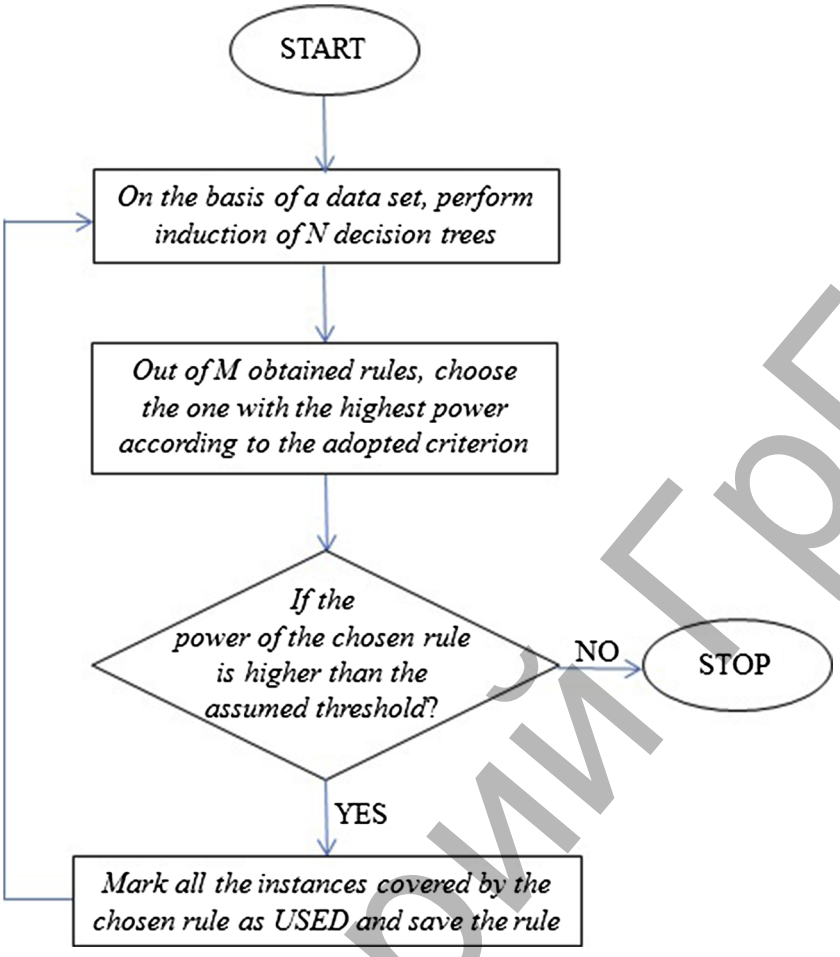

Fig. 2. Algorithm used to create a set of rules.

general, Random Forests enable to obtain better results than single trees. In order to make this advantage evident, certain variations among the generated trees are introduced. The variations can be ensured by random selection of a training set for each of the trees. Such a set contains a strictly defined percentage of the whole training set. Classification is performed similarly to other methods of combining classifiers. The most commonly chosen strategy is the majority vote.

2.2. Proposed method

In the present study, rules describing a data set according to the algorithm are shown in Fig. 2. Any criterion may be chosen for the assessment of the quality of rules; however, in the case when the induced rule covers the USED class, the power of this rule is determined on the basis of original classes. Owing to this, rules that cover instances belonging to the same class will not be eliminated as too ineffective, despite them having been covered be a previously used rule. If, however, the USED class is over half of the instances covered by a rule, then it is not taken under consideration as it will lead to overfitting. This last comment does not pertain to cases in which the number of already obtained rules is smaller than twice the number of classes.

The method that we used for the assessment of the quality (power) of rules was Laplace measure, described with the following equation:

$$Lapl = \frac{T+1}{N+K}$$

where:

T – number of instances correctly classified by a rule;

N – number of all instances covered by a rule;

K – number of classes.

It is worth noticing that this measure promotes not only the correctness of classification but also the rules that cover a larger number of instances. Owing to the use of Laplace measure, fewer rules should be obtained than in the case of other measures.

Using the induction algorithm in question may result in a rule being covered by two or more rules with a different conclusion (class). Owing to this, when rules attributing data to given classes are used, a priority consistent with the order of their creation must be attributed to individual rules. In this case, the analyzed instance will be attributed according to the rule with the lowest number (No.).

2.3. Experimental setup

2.3.1. The study group

The study was performed in a group of 27 children (12 PV and 15 children with typical foot as a control group). The characteristics of study subjects is presented in Table 1.

All participants and their parents were informed about the aim and course of the study. All parents of children involved in the investigation

Table 1
Characteristics of subjects.

Case	n	Number of used strides	Age ± SD [years]	BW ± SD [kg]
Typical	15	154	13.64 ± 1.91	50.8 ± 9.56
PV	12	73	12.60 ± 2.17	44.4 ± 11.46
AC	10	89	15.10 ± 1.52	54.3 ± 9.27

PV – pes planovalgus; AC – after correction; BW – body weight.

signed the necessary agreements and declarations before hospitalization, before surgery and before testing. The research has obtained the approval of the Ethics Committee at Grodno State Medical University, Belarus (approval number: 2011-8).

The measurements of gait were conducted in the Research Center of Resources-Saving Problems of the National Academy of Sciences of Belarus. The participants walked along a pathway in a comfortable and self-determined manner, thus, for a single participant we recorded many strides of human gait and the overall number of recorded strides was 360.

After the first round of investigation 10 children with PV were operated in the Grodno City Clinical Hospital of Emergency Care (Belarus). The surgery procedure is described below in Section 2.3.2. The children that have undergone the surgery (children After Correction - AC group) were re-examined. In total, 458 strides were recorded. In accordance with the results presented in a previously published study [15] as well as with the initial results obtained in the present study, 142 gait cycles were excluded from further analysis. These were the gait cycles that started and finished walking of the examined patients. These cycles were characterized by variable gait velocity typical for starts off and stops. For this reason, only 316 gait cycles were used in further analysis.

2.3.2. Surgery procedures

Ten children with PV were operated according to medical doctors suggestion. Two variants of surgical procedure were applied. The purpose of the surgical procedure No. 1 was lengthening of the outer lateral and dynamic stabilization of the medial column of the foot. The surgical procedure consists of the following steps:

Step 1 - Achilles tendon lengthening in the sagittal plane with medialisation of the insertion point. It's an open operation, performed by Z-tenotomy with excision of the lateral mass of the tendon of the heel bone or by the Hoke method.

Step 2 - transverse lengthening osteotomy of the anterior part of calcaneus using lever-type and lamellar distracters developed by us and implantation of a bone graft. The distracters allow constant visual control and do not cripple subtalar articulation on this operation stage exercise.

Step 3 – the tendon m. tibialis posterior is exposed and Z-shape crossed; tendon m. tibialis anterior is identified and mobilized for over 4–5 cm. Arthrotomy of talonavicular articulation with excision of the overdistensioned capsule and a part of spring ligament is performed on the lower-internal joint surface. On the lateral surface of the navicular bone the periosteum is incised and shifted, with the help of a chisel or oscillating saw the L-shaped bone groove is formed, where the tendon m. tibialis anterior is moved, which is fixed in this position by interrupted sutures and covered by periosteum on the top. The distal part of the tendon is fixed to the plantar surface of the medial cuneiform bone and to the base of the first metatarsal bone by interrupted sutures. Thus, new massive ligament is formed on the lower-internal part of the arch, and at the same time the abnormal supination of the forefoot is corrected.

Capsuloplasty of the talonavicular articulation is performed. Tendon m. tibialis posterior is sutured with shortening in the supine foot position.

In postoperative period plaster cast immobilization is made on the middle third of the thigh in knee flexion up to 30° and in normal (0°) position of the foot. Immobilization period is up to 8–10 weeks.

The surgical procedure No 2 is performed from the 4–5 cm transverse incision on the skinfold slightly above origin of the Achilles tendon. The heel-string is released for 4–5 cm upward, starting from its origin to calcaneal tuber. By the 5–7 cm vertical incision in the sagittal plane, this heel-string is divided into two equal parts. The outer portion of the heel-string is excised from the calcaneal tuber and it is stitched by two strong capron tendon sutures. On the internal surface of the upper edge of the heel bone a U-type transaction of the periosteum is made, which is exfoliated and moved downwards. The cortical plate 5 X 5 mm is removed and osseous groove 5–7 mm in depth is formed half of the outer portion of the Achilles tendon is moved inwards. On the floor of the osseous groove two transversal canals are formed by the bone drill (2 mm diameter) outside to the opposite cortical plate of the heel bone. Surgical suture is placed into these canals, which fixes the rotating part of the Achilles tendon. When sutures are tightened the heel-string goes in the osseous groove and sutures are tied up on the outer cortical plate. U-type flap of periosteum is stitched to displaced tendon. After the surgery is carried out, circular plaster cast immobilization is used for a period of 5–6 weeks.

2.3.3. Acquisition of gait

During the investigation the barometric insole of appropriate size was placed into the subject's shoes. Special shoes were used with a flat and firm sole to rule out a person's shoes feature influence. The tested children were allowed time for habituation to walk in the special shoes. The barometric insoles consisted of capacitive sensors (max. 240 sensors per insole), which allow to record with frequency of 300 Hz the distribution of pressure of the human plantar onto the contact surface while walking. The measuring range was 0.6–64 N/cm^2. The distance of walk was about 8 m. The tested children were allowed to start walking from their left or right foot. An effective walk with self-selected speed has been estimated visually or by the peaks on the pressure diagrams (Fig. 1). The times of immobility (static) and walk (dynamic) are shown on the pressure diagrams which enables to compare the left and right foot or steps stability.

There is a strong dependence of pressure's amplitude from a vertical projection of a walk's speed, therefore it is necessary to achieve not only uniformity and repeatability of a walk's character, but also no "jumping" gait of a patient.

2.3.4. Creation of dataset

In order to find the rules and then assess the ability of knowledge generalization by the created model the data was randomly divided into two disjoint sets: the training dataset and the test dataset. The training dataset consisted of 203 gait strides (98 typical, 48 PV and 57 AC) while the test dataset consisted of 113 strides (56 Typical, 25 PV and 32 AC).

The insole sensors were grouped into seven anatomic masks (*Z*) (Fig. 3):

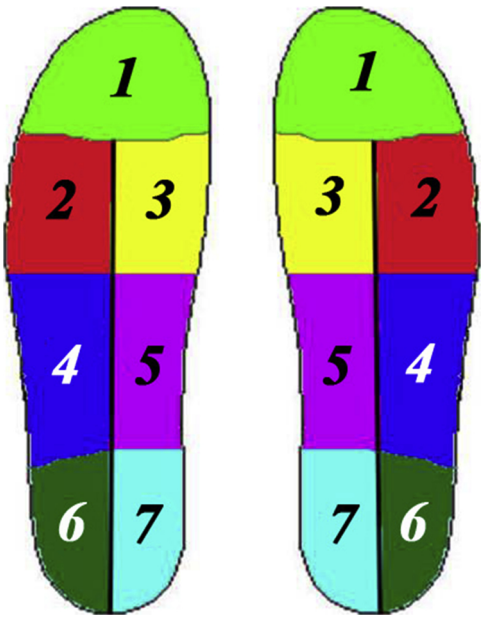

Fig. 3. The sensors insole divided into anatomic masks.

1 Mask 1 - toes,
2 Mask 2 - metatarsal heads (from 2nd to 5th),
3 Mask 3 - the head of 1 st metatarsal,
4 Mask 4 - the cuboid bone,
5 Mask 5 - the navicular bone,
6 Mask 6 - the lateral heel,
7 Mask 7 - the internal heel.

We divided the recorded time series into parts - the parts are the phases which are present in normal support phase of gait and they were:

- Loading response - which lasts 0–10% of gait cycle (*LR*),
- Mid Stance - which lasts 10%–30% of gait cycle (*MidSt*),
- Terminal Stance - which lasts 30%–50% of gait cycle (*TSt*),
- PreSwing - which lasts 50%–60% of gait cycle (*PreSw*).

The following parameters were calculated for every mask and gait phase:

- the average value of pressure normalized to body weight (*zn*),
- the time of i-th mask contact normalized to time of support phase (*t*),
- the percent of participation of the i-th mask in total foot loading (*%*).

Additionally, the first two parameters (*zn* and *t*) were determined for the whole foot during individual subphases of the stance phase.

As a result, we obtained a description of each of the gait cycles by means of 92 variables (4 subphases x 3 parameters x 7 masks + 4 subphases x 2 parameters = 92 variables). In the study, each of these variables was represented with an abbreviation consisting of three elements. The first element referred to the above described parameters: *zn*, *t* and '*%*'; the second part corresponded to the phase of gait (*LR*, *MidSt*, *TSt*, *PreSw*); the last element referred to anatomic masks of insole sensors (*Z1* – Mask1, *Z2* – Mask2, …, *Z7* – Mask 7, *Foot* – the whole insole without dividing into anatomic masks). As an example the abbreviation: '*znTst_Foot*' denotes the average value of pressure normalized to body weight during terminal stance in the whole foot (without dividing into anatomic masks), and '*%MidST_Z5*' denotes the percent of participation of mask 5 in the total foot loading during mid stance. The presented abbreviations are used in Tables 3–6.

3. Results and discussion

Most learning systems, including decision trees, tend to reproduce too closely the data in the training set. This results in a relatively poor ability to generalize the knowledge which is called overfitting. In order to limit this phenomenon, a 10-fold cross validation was used in this study when designing the decision trees. The Weka as well as Statistica software have been used to induce decision trees and to generate rules without decision trees (Table 3).

The rules induced by the respective trees vary in quality. A comprehensive, synthetic description of the ability of the obtained rules to describe a phenomenon is the percentage of steps from the training sequence correctly classified by the set of rules read from a given tree. These values, together with the percentage of correctly classified data from the training set and the number of rules and the number of uncovered instances, are presented in Tables 2 and 3. To compare the obtained results, classification of the obtained results using Random Forest was performed. An assumption was made that the classifier will consist of 100 base trees.

Analyzing the obtained values, it can be noticed that CART, CHAID, Exhaustive CHAID and C4.5 trees reached similar, relatively low values of correctly classified instances. The number of rules yielded by these trees is usually 10-11. An exception is the C4.5 tree, which created as much as 16 rules describing the dataset. As a result, overfitting was evident and the percentage of correct classifications of data from the test sequence in this case is by far the lowest at only 68.18%. In this view, two methods stand out in a positive way: Random Forest and the method proposed in this paper. The result yielded by Random Forest confirm that this is one of the best tree classifiers. In Random Forest, the number of rules for individual base trees is obviously different. In total, there are close to a thousand of these rules, which makes it impossible to use this method for interpretation of the obtained results, as mentioned earlier.

Our proposed method achieved a quality very similar to Random Forest. It should be emphasized, however, that in this case only instances covered by the used rules were used to calculate the percentage of correct classifications. If all uncovered instances were treated as classified incorrectly, then the results of classification would be 87.19% and 83.19% for the training and the sequence, respectively. It should be emphasized that such a treatment of uncovered instances is inappropriate as this would result in the proposed method indicating outlying measurements. A more detailed analysis of uncovered instances showed that as much as 17 of them from the training sequence and 9 from the test sequence are data describing the first or last steps of the tested patient. This indicates the need of analysis of only those gaits when the tested patient keeps constant stride speed.

The results presented in Table 3 show that the methods such as MODLEM which allow to generate rules without the need to build the decision tree have a little better accuracy than the methods showed in Table 2. Each of those methods, except from Decision Table, achieved 82–84% of correct classification rate (calculated on the testing set). However, in most cases those methods generated much more rules than the decision trees.

Tables 4–7 contain the rules obtained using the proposed method. The order of conditions in the individual rules is inconsequential as far as the result of the rule's action (classification) is concerned. However, during induction of trees, the conditions embedded in the rules initially ensure a division of input data into the most heterogeneous ones. This means that they contain parameters that vary the analyzed datasets to the largest extent. This is why these conditions should be paid special attention in the analysis.

A typical phenomenon is the decrease of Laplace measure with subsequent rules. This is largely the result of the decreasing number of instances covered by a given rule. It can be easily noticed that the number of rules is directly connected with the number of recognized classes. In the first case (Table 4), when there are 3 classes, the number of rules is almost twice as high as in Tables 5–7. The number of rules is also obviously connected with the ease of interpretation of the obtained results. Moreover, a higher data complexity results in the appearance of rules with multiple conditions (rule 1 and rule 4 in Table 4). Rules of this type are not very transparent and it seems that the analysis of pairs of specific cases would be a better solution.

When analyzing the individual rules, it is visible at first glance that one of the conditions differentiating patients with a typical foot from those with pes planovalgus (Tab. -5) is the higher value of normalized pressure in the Mid-Stance phase under the cuboid bone (*znMidSt_Z4*) for children with PV. This result may seem surprising at first, as pronation and flattening of the longitudinal arch during loading are commonly known and reported in the studies using measurements of pressures under the foot [21]. However, an analysis of average values of *znMidSt_Z4* for individual classes in the present paper (AC: 0.321 ± 0.175 N/(kg*cm2); PV: 0.509 ± 0.232 N/(kg*cm2); Typical: 0.212 ± 0.130 N/(kg*cm2)) indicates that this parameter is in fact a good differentiation between PV and Typical foot. The usefulness of *znMidSt_Z4* when differentiating between AC and Typical foot is obviously low and this is why it appears in Table 7. It is worth noting that in the previous study [22] the investigations were based on a quite similar group (Typical and PV) by means of a device of the same type (it used only 24 sensors as compared to our device with 256 sensors). The

Table 2
Quality of rules obtained from the tested decision trees.

Type of decision tree	Correct classifications (training set)	Correct classifications (testing set)	The numer of rules	The number of uncovered instances training/testing sets
CART	87.19%	78.34%	10	0
CHAID	85.71%	80.18%	11	0
Exhaustive CHAID	79.80%	76.96%	10	0
C4.5	83.74%	68.18%	16	0
Random Forest	95.71%	94.31%	several hundred	0
Proposed method	98.88%	94.06%	8	26/13

Table 3
Quality of rules obtained from the tested methods for rules extraction.

Type of method	Correct classifications (training set)	Correct classifications (testing set)	The numer of rules	The number of uncovered instances training/testing sets
Decision Table	80.78%	68.14%	21	0
JRiP	96.55%	82.30%	9	0
MODLEM	100%	84.96%	37	0
PART	98.03%	82.30%	16	0

Table 4
Rules obtained using the proposed method for the data describing AC, PV and Typical gait.

No.	Rules	Type of decision tree	Laplace measure	Covered patterns AC/PV/Typical
1.	IF (znTst_Foot < = 0.5205) AND (%MidST_Z5 < = 5.5465) AND (tPreSw_Z3 < = 0.1267) AND (tLR_Z4 < = 0.0933) AND (tTSt_Z6 < = 0.1633) AND (%MidSt_Z1 < = 5.6563) AND (%MidSt_Z7 > 13.1324) AND (znMidSt_Z6 < = 1.7511) AND (tMidSt_Z1 < = 0.1167) →"TYPICAL"	Ext. CHAID	0.9634	1/0/78
2.	IF (znTSt_Foot > 0.5205) →"PV"	C4.5	0.9048	0/18/0
3.	IF (znTSt_Z5 > 0.0058) AND (tPreSw_Z1 < = 0.1083) AND (tPreSw_Z4 > 0.01) AND (znTSt_Z2 < = 0.983) →"AC"	CHAID	0.8824	14/0/0
4.	IF (tPreSw_Z2 < = 0.1533) AND (tPreSw_Z2 > 0.1033) AND (znMidSt_Foot < = 0.4685) AND (%LR_Z2 < = 0.08173) AND (znMidSt_Foot > 0.2621) AND (znLR_Z6 < = 1.4623) AND (%MidSt_Z1 > 0.2505) AND (znPreSw_Z2 < = 0.6237) →"PV"	C4.5	0.9231	0/23/0
5.	IF (tPreSw_Z3 > 0.1533) →"AC"	C4.5	0.8667	12/0/0
6.	IF (tPreSw_Z1 > 0.1033) AND (tPreSw_Z1 < = 0.1367) AND (%LR_Z7 < = 33.0424) →"AC"	CHAID	0.8500	16/1/0
7.	IF (%PreSw_Z1 < = 45.3557) AND (%LR_Z5 < = 0.8727) AND (%PreSw_Z4 < = 8.2876) AND (%MidSt_Z1 > 3.4555) →"TYPICAL"	CART	0.8333	0/0/9
8.	IF (tPreSw_Z7 < = 0.0267) AND (tLR_Z3 < = 0.0367) AND (%TSt_Z1 > 26.0356) AND (znMidSt_Z4 > 0.312) AND (tLR_Z4 < = 0.11) →"PV"	C4.5	0.7500	0/5/0

PV – pes planovalgus; AC – after correction.

Table 5
Rules obtained using the proposed method for the data describing PV and Typical gait.

No.	Rules	Type of decision tree	Laplace measure	Covered patterns PV/Typical
1.	IF (znMidSt_Z4 < = 0.4028) AND (tTSt_Z6 < = 0.1633) AND (%TSt_Z1 < = 30.6434) →"TYPICAL"	C4.5	0.9780	1/88
2.	IF (znMidSt_Z4 > 0.4228) AND (tPreSw_Foot > 0.11) →"PV"	Exst. CHAID	0.9730	35/0
3.	IF (znLR_Z7 < = 0.8064) →" PV"	CART	0.9167	10/0
4.	IF (znMidSt_Z2 > 0.0071) AND (tMidSt_Z3 < = 0.075) AND (znLR_Z5 > 0.0027) →"TYPICAL"	CART	0.9000	0/8

PV – pes planovalgus.

Table 6
Rules obtained using the proposed method for the data describing AC and Typical gait.

No.	Rules	Type of decision tree	Laplace measure	Covered patterns AC/Typical
1.	IF (tTSt_Z4 < = 0.2033) AND (tPreSw_Z5 = 0) →" TYPICAL"	CHAID	0.9683	1/60
2.	IF (tTSt_Z4 > 0.2) AND (znLR_Z7 < = 0.86) AND (%TSt_Z4 > 5.0734) →"AC"	C4.5	0.9783	44/0
3.	IF (tPreSw_Z7 < = 0.02) AND (tTSt_Z4 > 0.2033) AND (tTSt_Z4 < = 0.2333) AND (znLR_Z7 > 0.9182) AND (tLR_Z4 < = 0.09) →"TYPICAL"	CHAID	0.9615	0/24
4.	IF (znLR_Foot > 0.3349) →"AC"	CART	0.9167	10/0
5.	IF (tTSt_Z2 > 0.2417) AND (znLR_Foot > 0.1961) →"TYPICAL"	CART	0.9333	0/13

AC – after correction.

Table 7
Rules obtained using the proposed method for the data describing AC and PV gait.

No.	Rules	Type of decision tree	Laplace measure	Covered patterns AC/PV
1.	IF (znPreSw_Z1 < = 0.6868) AND (%MidSt_Z5 < = 3.0445) AND (znPreSw_Z4 < = 0.1843) →"AC"	CART	0.9623	50/1
2.	IF (%MidSt_Z5 > 3.045) AND (znTSt_Z6 < = 0.0837) →"PV"	CART	0.9545	0/20
3.	IF (%MidSt_Z5 < = 3.0319) AND (znPreSw_Z1 > 0.6708) →" PV"	CART	0.9474	0/17
4.	IF (znPreSw_Z4 > 0.1727) →"PV"	C4.5	0.8889	0/7
5.	IF (tMidSt_Z3 < = 0.2483) AND (%MidSt_Z5 > 2.9554) AND (%MidSt_Z4 < = 17.8349) →"AC"	CHAID	0.8571	5/0

PV – pes planovalgus; AC – after correction.

insole sensors were grouped into only two anatomic masks (lateral and distal part of the foot). The reported results show that children with typical feet have bigger pressure on lateral part of the foot a little more often than children with pes planovalgus. The second important parameter in this case is *tPreSw_Foot*, whose values are 0.126 ± 0.013 s for PV and 0.110 ± 0.234 s for Typical foot. After considering the second condition of this rule (*znMidSt_Z4* > 0.4228), the difference between the average values of this parameter increased and was 0.128 for PV and 0.106 for Typical group.

The expected differentiation of pressure on foot in mask *Z5* during the Mid-Stance phase is used to discriminate between persons belonging to groups PV and AC (Table 7). Parameter *%MidSt_Z5* plays an especially important role here as it enables to correctly classify almost 90% of all AC cases. The values of this parameter are 6.980 ± 9.802% for PV and 1.401 ± 1.266% for AC. The high variability of the parameter's values for PV resulted in the appearance of rule 3, which for PV instances with a low *%MidSt_Z5* value also uses parameter *znPreSw_Z1*, whose value for instances meeting condition *%MidSt_Z5* < = 3.0319 is 0.348 ± 0.177% for AC and 0.771 ± 0.368% for PV.

Table 5 presents the rules that indicate the differences between typical gait and gait following a surgery. In this case the greatest difference can be observed in mask 4 during terminal stance. In the first two rules parameter *tTSt_Z4* plays the main role. Its values, i.e. 0.254 ± 0.044 s for AC and 0.184 ± 0.051 s for Typical foot indicate a significantly longer duration of this subphase, which is characteristic of a relatively large number of feet pathologies [23]. It is worth adding that for children with PV, this time, at 0.233 ± 0.040, is very close to AC time. Parameter *%TSt_Z4* is 16.677 ± 7.141% for AC and 7.862 ± 5.287% for Typical (14.956 ± 7.131% for PV), which as before indicates disorders of foot transition.

It is quite interesting to compare the rules induced by means of the single decision trees (Table 2) to the rules obtained using the proposed method. In case of CART and CHAID the most important parameter which divides the data into two the most homogeneous subsets is *zn_MidSt_Z4*. It should be underlined that the biggest differences in data exit for Typical and PV groups, so *zn_MidSt_Z4* is the attribute which has been indicated in Table 5, too. The different way for choosing the attributes by CART and CHAID results with a slightly different value of threshold (0.4059 for CART and 0.3969 for CHAID). However the next important attributes are different for those decision trees: *tTSt_Z4* and *%MidSt_Z1* for CART and *tTSt_Z6* and *znMidSt_Z1* for CHAID. In the case of Exhaustive CHAID and C4.5 the most important attribute is *znTst_Foot* (see at rules no 1 and 2 in Table 4). It is worth noting that the threshold value is the same for those decision trees.

In the study by Ledoux et al. [23] the authors showed that in the case of adults with pes planovalgus, there is a statistically greater plantar pressure under 1st metatarsus than in persons with a typical foot. On the contrary, Pauk et al. [21] observed a lower pressure distribution in the metatarsal heads compared to children with typical foot. The results obtained in the present study show that pressure under 1st metatarsus (mask *Z3*) is not at all considered as a condition of any of the rules. Obviously, this does not prove that there are no differences between the pressures in this area; this rather means that other parameters differentiate the tested cases better.

4. Conclusions

Measurement of ground pressure during gait is useful for the assessment of foot and gait pathologies. The presented method for mining rules useful in this respect may be successfully used by persons with a typically medical knowledge and could improve the understanding of the human gait phenomenon. The proposed method focuses on mining the relationships between data the most typical for the tested groups of patients. Data that are atypical measurements in any way are practically disregarded when creating rules. There is obviously no reason why this method could not be used in the case of other data as well.

Further work in this area should focus on simplifying rules and limiting their number. The first issue could probably be realized by using the so-called cross trees, which test more than one parameter in each node. Interpretation of rules obtained in this manner is an open matter. The number of rules could be limited using a more rigorous stop criterion (e.g. Laplace measure > 0.85). However, the impact of such actions on the quality of the obtained results should be tested.

Conflict of interests

The authors declare no conflict of interests.

Financial disclosure

This work was financed by the Polish Ministry of Science and Higher Education within the framework of the project no. S/WM/1/2017.

The author contribution

Study Design: Marcin Derlatka, Mikhail Ihnatouski, Marek Jałbrzykowski, Vladimir Lashkovski.

Data Collection: Mikhail Ihnatouski.

Statistical Analysis: Marcin Derlatka, Marek Jałbrzykowski, Łukasz Minarowski.

Data Interpretation: Marcin Derlatka, Mikhail Ihnatouski, Vladimir Lashkovski.

Manuscript Preparation: Marcin Derlatka, Mikhail Ihnatouski, Marek Jałbrzykowski, Łukasz Minarowski.

Literature Search: Marcin Derlatka.

Funds Collection: Marcin Derlatka.

References

[1] Chen JP, Chung MJ, Wang MJ. Flatfoot prevalence and foot dimensions of 5–to 13-year-old children in Taiwan. Foot Ankle Int 2009;30(4):326–32.
[2] Das SP, Das PB, Ganesh S, Sahu MC. Effectiveness of surgically treated symptomatic plano-valgus deformity by the calcaneo stop procedure according to radiological, functional and gait parameters. J Taibah Univ Med Sci 2017;12(2):102–9.
[3] Jafarnezhadgero AA, Shad MM, Majlesi M. Effect of foot orthoses on the medial longitudinal arch in children with flexible flatfoot deformity: a three-dimensional moment analysis. Gait Posture 2017;55:75–80.
[4] Saraswat P, MacWilliams BA, Davis RB, D'Astous JL. Kinematics and kinetics of normal and planovalgus feet during walking. Gait Posture 2014;39(1):339–45.
[5] Derlatka M, Ihnatouski M. Decision tree approach to rules extraction for human gait analysis. Artificial intelligence and soft computing. Berlin/Heidelberg: Springer; 2010. p. 597–604.

[6] Rosenbaum D. Assessing pediatric foot deformities by pedobarography. Springer International Publishing; 2016.
[7] Barn R, Waaijman R, Nollet F, Woodburn J, Bus SA. Predictors of barefoot plantar pressure during walking in patients with diabetes, peripheral neuropathy and a history of ulceration. PLoS One 2015;10(2):e0117443.
[8] Deepashini H, Omar B, Paungmali A, Amaramalar N, Ohnmar H, Leonard J. An insight into the plantar pressure distribution of the foot in clinical practice: narrative review. Polish Ann Med 2014;21(1):51–6.
[9] Fernández-Seguín LM, Mancha JAD, Rodríguez RS, Martínez EE, Martín BG, Ortega JR. Comparison of plantar pressures and contact area between normal and cavus foot. Gait Posture 2014;39(2):789–92.
[10] Miljkovic D, Aleksovski D, Podpečan V, Lavrač N, Malle B, Holzinger A. Machine learning and data mining methods for managing Parkinson's disease. Machine Learning for Health Informatics. Springer International Publishing; 2016. p. 209–20.
[11] Derlatka M, Pauk J. Data mining in analysis of biomechanical signals. Solid State Phenom 2009;147:588–93.
[12] Rudek M, Silva NM, Steinmetz JP, Jahnen A. A data-mining based method for the gait pattern analysis. Facta Universitatis. Series: Mech Eng 2015;13(3):205–15.
[13] Armand S, Watelain E, Roux E, Mercier M, Lepoutre FX. Linking clinical measurements and kinematic gait patterns of toe-walking using fuzzy decision trees. Gait Posture 2007;25(3):475–84.
[14] Jones GG, Kotti M, Wiik AV, Collins R, Brevadt MJ, Strachan RK, et al. Gait comparison of unicompartmental and total knee arthroplasties with healthy controls. Bone Joint J 2016;98(10 Supple. B):16.
[15] Derlatka M, Bogdan M. Combining homogeneous base classifiers to improve the accuracy of biometric systems based on ground reaction forces. J Med Imaging Health Inform 2015;5(8):1674–9.
[16] Friedman JH, Popescu BE. Predictive learning via rule ensembles. Ann Appl Stat 2008:916–54.
[17] Deng H. Interpreting tree ensembles with intrees arXiv preprint arXiv:1408.5456 2014.
[18] Hara S, Hayashi K. Making tree ensembles interpretable arXiv preprint arXiv:1606.05390 2016.
[19] Rajput A, Aharwal RP, Dubey M, Saxena SP, Raghuvanshi M. J48 and JRIP rules for e-governance data. Int J Comput Sci Secur (IJCSS) 2011;5(2):201–7.
[20] Stefanowski J. On combined classifiers, rule induction and rough sets. Transactions on rough sets VI. Berlin, Heidelberg: Springer; 2007. p. 329–50.
[21] Pauk J, Daunoraviciene K, Ihnatouski M, Griskevicius J, Raso JV. Analysis of the plantar pressure distribution in children with foot deformities. Acta Bioeng Biomech 2010;12(1):29–34.
[22] Marmysh AG. Plantar pressure distribution in children with pes planovalgus (Особенности распределения подошвенного давления при плоско-вальгусной деформации стопы у детей). J Grodno State Med Univ 2017;15(4):400–4.
[23] Ledoux WR, Hillstrom HJ. The distributed plantar vertical force of neutrally aligned and pes planus feet. Gait Posture 2002;15(1):1–9.